

Data Provenance and its Implications for the Development of E-Science Systems

Kate Ho

January 30, 2006

Introduction

- ▶ 2nd Year PhD student supervised by Rob Procter, Stuart Anderson and Mark Hartswood
- ▶ Part of the Social Informatics Cluster (SIC)
- ▶ Funded by the DIRC project (Dependability in Computer based Systems)
- ▶ This talk is based on my PhD thesis - which is about the **requirements capture process in E-Science Systems**

Contents

- ▶ Introduction
 - ▶ What is E-Science?
 - ▶ What is Data Provenance and why is it important?
- ▶ Research Approach
- ▶ Empirical Data
- ▶ Future Work

Data Provenance within E-Science Systems

This talk is on...

What impact will data provenance have on E-Science systems?

Data Provenance within E-Science Systems

very simply...

What impact does data sharing, production and analysis have on the building of E-Science Systems?

Introduction to E-Science

"Each of the [e-science] projects had great aspirations for bringing about serious change in the world of science and engineering by facilitating new approaches to data analysis, the extraction of knowledge from very disparate sources, and the application to experiments that were until now unimaginable. The future of large-scale, collaborative science is here now ..."

EPSRC report on E-Science (2004)

"The UK E-Science programme has initiated significant developments that allow networked grid technology to be used to form virtual co-laboratories"

Atkinson et al. (2002)

E-Science

- ▶ Regarded as a new configuration of doing science
- ▶ E-Science is ...
 - ▶ information intensive
 - ▶ large scale
 - ▶ distributed
- ▶ E-Science crosses international, inter-disciplinary and inter-group research
- ▶ Example - E-DiaMoND and AstroGRID

E-Science

- ▶ The technologies used to enable the e-science infrastructure are Grid Technologies and Service Orientated Architectures (S.O.A.s).
- ▶ Grid technologies and S.O.A.s provide the protocols and tools which allow large amounts of data to be shared between heterogeneous resources.
- ▶ It also allows large amounts of computation power to be accessed on demand.
- ▶ Results in an infrastructure which has data on demand, ways of allowing scientists to communicate (e.g. Access Grid meetings), new ways of analysing and comparing data (grid based tools).

What is Data Provenance?

- ▶ We use the term **Data Provenance** to describe the process in which a piece of data is produced, stored, shared and analysed.
- ▶ In other words, it is the history of a piece of data

The History of a piece of data!?!

- ▶ In "traditional" science, each scientist designs, runs and records their own experiments. They are usually there to observe and make observations about the experiments that they run
- ▶ Most importantly, this means that the scientist will understand the **contingencies** that the data contains
- ▶ If they aren't satisfied with the way that the experiment was conducted, they would repeat the experiment again
- ▶ Essentially, the scientist knows **where the data was produced, under what conditions it was produced under, how it is recorded and what the data is analysed afterwards**

Why is Data Provenance different in E-Science Systems?

Under the vision of e-science, there is the promise of **flexible** data. Flexibility can be:

- ▶ Data being available on demand - anywhere, anytime, any scientist
- ▶ Ability to repurpose data
- ▶ For example, Combe-Chem at Southampton University

Because of this flexibility, e-scientists can move onto different research problems quickly

What is the problem of flexible data?

- ▶ However, in traditional science, there is a need for consistency in the way that data is collected and/or analysed.
- ▶ Otherwise, you can't compare the data as they are not like-for-like.
- ▶ For example, tissue preparation in medical research.

Research Approach

- ▶ The research approach was a **10 month observational study** of the development of an e-science system
- ▶ Development of the IT infrastructure to support large scale translational research project
- ▶ An e-science project that the Social Informatics Cluster is involved in
- ▶ As a result, it was easy to gain access

Background to case study: TransProject

- ▶ The aim of TransProject is to build an infrastructure that facilitates the recruitment of patients into cancer research
- ▶ TransProject is part of network of translational cancer research centres across the UK
- ▶ The system's scope spans from supporting nurses obtaining consent from the patients through to human tissue being analysed
- ▶ Has to link up and deal with several different data sources such as the national cancer registry

Project Team

- ▶ Principal Investigators - A board of 11 P.I.s, each in various different fields including oncology, genetics, informatics
- ▶ Research (Project) Manager
- ▶ Database Manager
- ▶ E-scientist (employed by Informatics)
- ▶ E-scientist (employed by genetics department)
- ▶ 2 Research Nurses
- ▶ Lab Technician

Examples from Empirical Work

I am going to present 3 examples from the case study of how data provenance has impacted on the development of an E-Science System:

- ▶ Example 1: Standard Operating Procedures
- ▶ Example 2: Ethics Approval
- ▶ Example 3: Decoupling of Data Production and Analysis

Example 1: Standard Operating Procedures (S.O.P.s)

- ▶ In this research project, they will take tissue and blood from cancer patients and input them into a database
- ▶ To do so, they have to follow Standard Operating Procedures (S.O.P.s)
- ▶ These S.O.P.s are rules by which work processes have to be followed - whether these are for consistency reasons (so that all samples are treated the same) or whether these are for health and safety reasons
- ▶ The S.O.P.s stipulate what how the data is collected and processed
- ▶ The developers have said that they need to know what the working processes of the system are from the very start

Example 1: Standard Operating Procedures (S.O.P.s)

- ▶ Two major points here:
 - ▶ 1. Because if you change the processes mid-way, then they can't easily compare the data as they are not like-for-like. And they didn't like having inconsistencies in their database
 - ▶ 2. In traditional systems development, it is expected that working practices and processes in a system evolve and adapt. However, that flexibility does not exist here.

Example 2: Ethics Requirements

- ▶ In medical research, ethic approvals needs to be gained before any work can be carried out with patient data
- ▶ To gain ethics approval you have to detail all working processes, all S.O.P.s, all the documents used and how the data is stored afterwards
- ▶ Any changes made in dealing with patients or how the data is used/stored has to gain approval from the ethics committee - which only meets once per quarter
- ▶ In this case, there is a high cost associated with any changes made to processes
- ▶ Its all about data consistency

Flexibility vs Stability?

Flexibility

- ▶ Quick to move into new directions of research
- ▶ Collection of data from different resources
- ▶ Use of data to address future research questions/topics
- ▶ Can collect more data and hence provide more novel service

Stability

- ▶ Consistency in the processes in order to compare like-for-like data
- ▶ Easier to build systems which will not change
- ▶ Set/pre-defined processes will result in better systems being built
- ▶ Need to know how the data is stored and treated on a long term basis

Example 3: De-coupling of data production and analysis

- ▶ E-Science creates a separation between data production and utilisation
- ▶ For example, in a related cancer study, one of the workers commented about the "age of onset" field
- ▶ How do you define the age of onset (i.e. diagnosis?) Is it when the person is told about the cancer? Is it when the doctor first recognises it as a cancer? Sometimes, doctors do further testing in order to "break in" the news to the patient that they have cancer.
- ▶ Which one do you take?
- ▶ Its easy if you have a commonly, locally shared understanding of a term (e.g. individually or within a group) but not when the data has to be shared and distributed across a number of research groups.

Implications for design

- ▶ Data is incredibly important in the context of scientific systems
- ▶ The understanding of context of data is key to being able to realise the E-Science vision
- ▶ There is a need to support the "work" of science - we build computer systems to support other people's work (e.g. call centres) - this is no different, we're just building systems to support scientists' work
- ▶ There might be a possible solution in realising the flexible data vision

The Immediate Remedy ... Meta Data

- ▶ Meta Data is data that describes how a piece of data is formed - i.e. provide its context
- ▶ For example, a piece of meta data could describe what the experimental conditions were, and when the sample was produced or analysed etc.
- ▶ Using meta data, it is hoped that data can be re-purposed easily. After all, knowing what conditions it was formed under means that you can understand how it differs from another piece of data, and therefore, compensate for its differences

The Immediate Remedy ... Meta Data - the Flaw

- ▶ However, the biggest problem for meta data is (like all standardisation processes):
 - ▶ "What meta data gets stored?"
 - ▶ "Who gets to decide what meta data gets stored?"
 - ▶ "Who decides what comparisons are made and how?"
 - ▶ "How much should you collect?"
 - ▶ "Who is going to collect it?" (or rather, how to collect it without creating extra work)

Social Informatics Cluster ... a brief intro

- ▶ The Social Informatics Cluster (SIC)
- ▶ Rob Procter, Mark Hartswood, Roger Slack, Alex Voss, Jenny Ure, Conrad Hughes (and me)
- ▶ Main research interests in the group are e-science, e-health, system development methodologies, ethnomethodology
- ▶ Based at 1 Buccleuch Place

Thank you for listening! Any questions?